UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

DOES GENE EXPRESSION VARIABILITY CHARACTERIZE CELL AND TISSUE HETEROGENEITY?

Máster Universitario en Bioinformática y Biología Computacional

Autor: PAZ CABEZAS, Mateo

Tutora: SÁNCHEZ CABO, Fátima
Co-Tutor: WERE, Felipe

Unidad de Bioinformática
Centro Nacional de Investigaciones Cardiovasculares (CNIC)

Septiembre, 2019

ABSTRACT 2
INTRODUCTION 3
Variability in Gene Expression 3
Statistics for dispersion measurement 3
Considerations for variability measurements 4
Immune System 5
Lymphocyte lineages 6
Monocytes (CD14) 7
Lymphoid Cells 7
T Helper Cells (CD4) 7
Citotoxic T Cells (CD8) 8
B cells (CD19) 8
Natural Killer (CD56) 9
Lymphoid Progenitor Cells (CD34+) 9
Methodologies to identify VEGs 9
scVEGs 10
Seurat 10
OBJECTIVES 12
MATERIAL AND METHODS 12
Datasets 12
Bioinformatic procedures 13
R session info: 13
SC data processing 13
Bulk Data processing 13
Seurat VEG procedures 13
scVEGS 13
Gene Set Enrichment analysis 14
Statistical Analysis 14
RESULTS 15
Comparing methodologies 15
Single cell 15
Bulk 16
Comparing variability profiles 18
Biological characterization through Variability in SC 23
Gene Set Enrichment Analysis 23
Biological features enrichment 25
Biological characterization through Variability in Bulk 26
DISCUSSION 29
Methodologies 29
Variability comparison 29
SEURAT::VST SC 29
SEURAT::VST Bulk 30
Future approaches 30
BIBLIOGRAPHY 32

ABSTRACT

Gene variability is described as the amount of dispersion in the expression values of a gene within a determined dataset. This value have been studied to determine genes that are affected by high environmental regulation, identify transcriptional regulators. But may the study of highly variable genes characterize different cell populations or conditions as mean expression values does? In this present work, we have observed that in Single cell expression datasets, variable expressed genes analysis (using Seurat::VST procedure) allow us to identify unique pathways, functions and cell locations that, characterizing different cell types, degrees of relationship and differentiate homogeneous from heterogeneous datasets.

INTRODUCTION

Variability in Gene Expression

Variability is described as the amount of dispersion in a given distribution, in the case of a biological analysis of gene expression, it would be considered as differences in expression within the same condition or “expression noise”.

During the last decade several works have tried to assess whether gene expression variability can provide insights on different cell populations beyond those acquired from standard analysis of gene expression means [1][2].

Analysis of gene variability within groups that are seemingly homogenous have so far been valuable to predict the outcome of diseases[3], shaping innate immunity [4] or identifying transcriptional regulators in the development of early human embryos[5]. In addition, identification of variably expressed genes (VEGs) can suggest disruptions or dysregulations of biological processes[1] .

In this scenario, single cell RNA sequencing (scRNA-Seq) has opened new avenues to study gene variability in clonal subpopulations of a determined sample. Different mechanisms contribute to the variability in homogenous cell populations, ranging from inherent variations in the biochemical process of gene expression itself, intrinsic and extrinsic noise and phenotypic plasticity of cells due to variations in the local micro-environment[1].

Compared to classic bulk RNA-seq, data obtained with scRNA-Seq techniques has some specifics due to the low mRNA content available from each individual cell analyzed and the variability derived from phenotypic plasticity and technical bias. Given that, scRNA-Seq data features lower read counts, a higher amount of outliers[6] and an increased inner variability between cells sharing the same conditions, being this variability more prominent in highly expressed genes[7]. These features result in a multimodal distribution for the expression values obtained[8].

This bimodal distribution is constructed as a mixture of Poisson distributions, where gene expression is modelled as a stochastic Poisson process depending on the RNA synthesis (influenced by the promotor of the gene and the upstream regulators), and RNA degradation, being both described as a Poisson process. Considering this RNA birth-death becomes a doubly stochastic (mixed) Poisson process[9]. The negative binomial distribution represents a good intersection between computational complexity and biological simplicity as NB distribution is defined through only two parameters

Statistics for dispersion measurement

Three main statistics might be used to address variability, i.e. SD, CV and MAD, and all of them share some conceptual similarities in their mathematical definition (Figure 1)[10].

SD: Is dependent on the value of average expression, but simulations have shown a lower correlation with average expression than previously expected [11]
CV: Is affected by zero-inflation or near zero levels.

MAD: Is a robust measurement of dispersion that behaves well in presence of outlier data points.

A consensus on which estimator should be adopted to address variability have not been achieved yet, and performance is still data-specific [12]. SD and CV are easier to interpret as they are in the same dimensions as mean. In any of the cases, variability estimators are **highly correlated with the average expression**, then **trends observed** for expression variability may simply be **recapitulated by those observed for average expression**. In an ideal scenario, the estimator of gene expression variability should be uncorrelated with average gene expression

Let $x_1, \ldots, x_n$ denote a univariate dataset, then

$$\mathrm{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\mathrm{CV} = \frac{\mathrm{SD}}{\bar{x}}$$

$$\mathrm{MAD} = median(|x_i - median(x)|).$$

Figure 1: Main statistics used for variability determination for a given x dataset[10].

Considerations for variability measurements

Standard error of methods to calculate over-dispersion tend to be larger than those designed to address differences by mean estimates[12], causing that statistical analyses based on average expression as DEG analysis [13]show higher reproducibility.

Given that, and putting aside the effect of technical variability, there are some general considerations that must be taken into account in order to perform variability studies [11].

Sufficient amount of samples: Technical variability has a higher impact on variability studies than in mean expression analysis, therefore a significantly higher amount of samples is required

Reducing batch effect: As technical bias adds a confusion effect to the variability measurement that can mask the biological dispersion of gene expression, is important to reduce its impact as much as possible. This can be achieved reducing batch effect distributing samples evenly through different batches, and also performing technical replicates.

Immune System

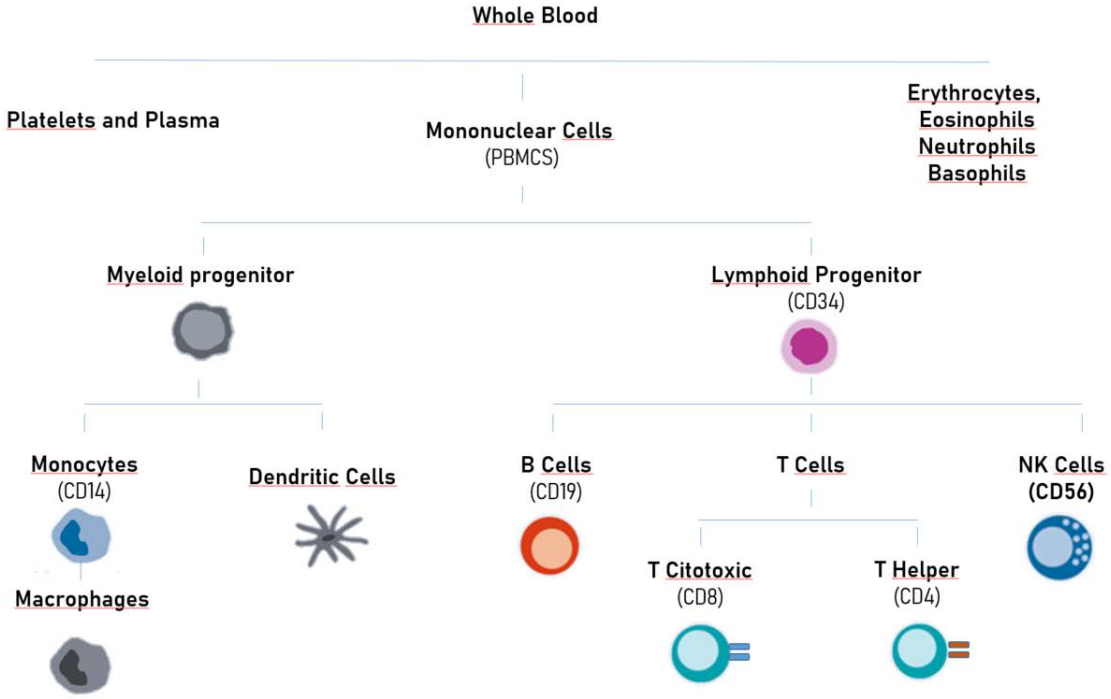

Figure 2: Scheme of the Cell components of whole blood and different lineages within PBMCS. Cells analyzed in the present manuscript are highlighted in colour and marked with the cluster of differentiation (CD) used for it selection.

Peripheral blood mononuclear cells (PBMCs) are blood cells featuring a round nucleus, including monocytes, lymphocytes, dendrytic cells and natural killer cells (Figure 2). Isolation of PBMCs performed by centrifugation of peripheral blood supplemented with anticoagulants in a density gradient using a hydrophilic colloid[14].

Surface markers expressed by immune cells are referred as the Cluster of differentiation nomenclature (CD), a protocol stablished in 1991 to standardize the designation and identification of cell phenotypes. Even though a particular CD is not specific of a single cell type or even a whole lineage, when combined they are a very powerful tool for immune cells characterization. Even though the complete list of determinants features more than 350 molecules with specific functions such as receptors for ligands and further signaling, cell adhesion, etc. Few of them are more widely spread for immune cells identification and isolation using flow cytometry procedures[15].

Cell Type	%	Frequencies ($x10^3$ Cells, Isolated from 1 mL Blood)
T cells ($CD3^+$)	60	300–1200
T helper cells ($CD3^+$, $CD4^+$)	70 of T cells	210–840
Cytotoxic T cells ($CD3^+$, $CD8^+$)	30 of T cells	90–360
Monocytes/macrophages ($CD14^+$)	15	75–300
B cells ($CD22^+$)	10	50–200
Natural killer (NK) cells ($CD56^+$/$CD16^+$)	15	75–300

Table 1: Cell components of PBMCs in percentage and frequencies[16].

The adaptive immune response is subdivided into functional groups representing humoral and cellular immunity, based on participation of the two major cell types. Humoral immunity involves B lymphocytes (also called B cells) which synthesize and secrete antibodies. Cellular immunity involves effector T lymphocytes(also called T cells) which secrete immune regulatory factors following interaction with specialized processing cells (called antigen presenting cells; APCs) that show the lymphocytes foreign material in the context of self-molecules[16].

Lymphocytes continuously leave the blood vessels, migrating throughout the body where they perform surveillance activities

Lymphocyte lineages

Leukocyte is the term given to any white blood cells that play a functional role in the immune response. They can be classified into two main groups depending on the developmental path taken during development in the bone marrow.

Myeloid cells represent the first line of defense as the most important part of the innate immunity. They provide a rapid response (in a range of minutes to hours) against pathogens in a nonspecific way, recognizing structural motifs and patters in molecules and secreting soluble activators and proinflammatory mediators, those mechanism involves no memory upon a future exposure[15]. Cells from myeloid lineage includes dendritic cells (involved in quick generic response to pathogens); phagocytic leukocytes as motile neutrophils, and monocytes (the circulating precursor of macrophages); eosinophils basophils and mast cells (in charge of parasite defense and allergic chain reactions).

In the other hand, lymphoid cell types are involved in specific immunity. Specific immunity creates a specific response to pathogens, creating a memory that will allow the organism to develop a quicker response to a future exposure, mediated by antibodies and cytokines. Lymphoid cells have receptors to recognize and physically

interact with antigens, B lymphocytes and the T lymphocytes. A functionally related set of cells considered also as lymphoid are the natural killer cells(NK cells).

In this manuscript we will focus on those cells selected by the following CDs: CD14 (Monocytes), CD19 (B lymphocytes), CD4 (T Helper lymphocytes, CD56 (NKs), CD8 (Cytotoxic T cells) and CD 34 (Hematopoietic precursors) (Table 2 and Figure 2).

Monocytes (CD14)

Monocytes are the largest of the leukocytes, they belong to the myeloid lineage, along with macrophages and dendritic cells. They are produced in the bone marrow before and the lifespan of a circulating monocyte is fairly brief and most undergo apoptosis after about 24 h[17].

Monocytes develop into macrophages after leaving the cell circulation in different tissues, performing surveillance functions against pathogens or eliminating dead cells. Macrophages are able to detect products of bacteria and other microorganisms using different receptors as Toll-like receptors (TLRs).The can also function as APCs to T cells and release cytokines to trigger and modulate immune cells (as seen in Helper T cells)[18].

Lymphoid Cells

Lymphoid cells are the principal cells in charge of the acquired immunity, responding to pathogens (microbes and pathogens) after myeloid cells mediation[16].

Lymphoid cells are separated in three lineages, the B Cells, the T cells and the NK cells.

T cells

T cells are produced in the thymus (where the T comes from) and mediate between cellular and humoral immunity. T Cells recognize antigens through their T cell receptor (TCR)[19]. But they cannot bind antigen directly, it needs to have processed by an antigen presenting cell (APC) and presented through the major histocompatibility complexes (MHC), a molecule present in APCs cell surface. Despite of the similarity with the antibody, TCR is only present on its surface and is not secreted when T cells are activated, and the TCR-MHC binding is unstable, requiring co-receptors[20]. This presentation process requires a constant migracion of T cells to secondary lymphoid organs to encounter APCs.

T Helper Cells (CD4)

Helper T cells interact with the different mediators of adaptative immune response shaping their activity: Stimulating antibody production of B cells, activating Macrophages through IFNg[21], Citotoxic T Cells and recruiting antigen presenting cells. They also secret cytokynes that can act on epithelial and smooth muscle cells.

CD4+T cells recognise peptides presented on MHC class II molecules, from antigen presenting cells (APCs), depending on the strength of TCR signaling and the cytokine microenvironment (Figure 3), activated T cells differentiate into distinct Th lineages[22]. Most of the modulations that T helper cells perform in the immune system (Figure 3) relies on the production of diverse cytokines to modulate the innate immunity, this process is tightly transcriptionally controlled[23].

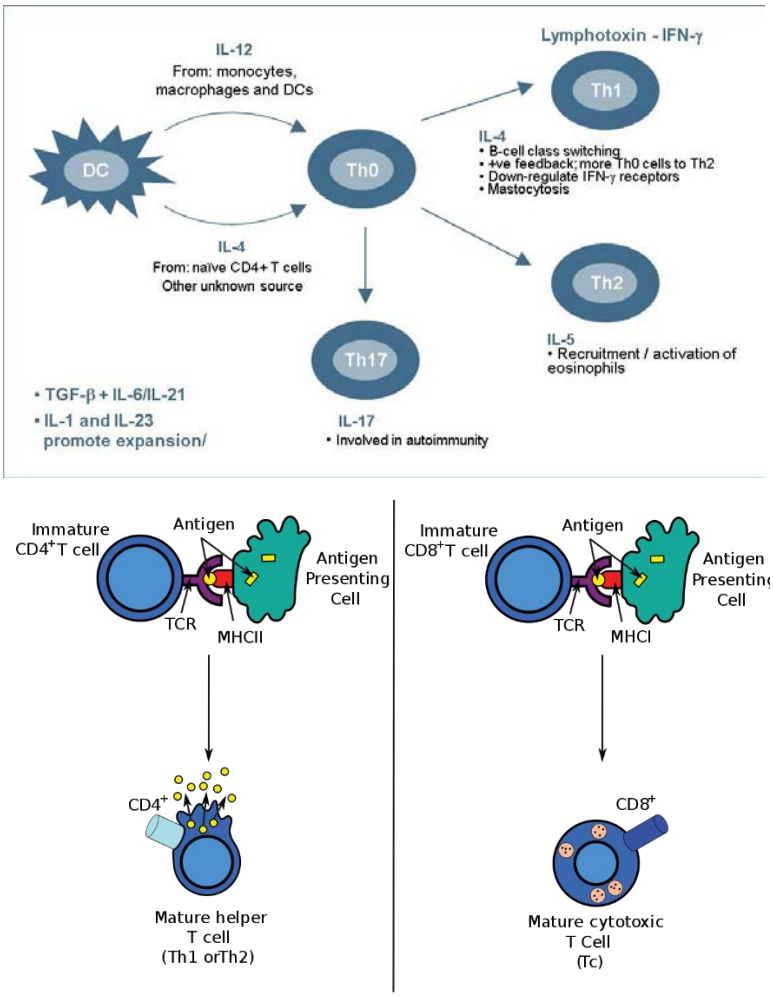

Figure 3: Different Interleukins involved in Helper T cell differentiation. In this process mediated by antigen presenter cells the cytokine microenvironment determinates the T cell fate[22]*. On the bottom, antigen presentation procedure between APCs and T cells, mediate by TCR and MHCI/MHCII.*

Citotoxic T Cells (CD8)

Citotoxic T Cells can induce apoptosis in malignant tumoral cells, or those cells affected by intracellular pathogens; this defense against abnormal "self-cells" (so called cellular immunity) is mediated by the liberation of cytolytic granules. This vesicles contain pore-forming proteins, proteases, granulolysins (which degrade membrane lipids) and ligands for membrane death receptors (as CD95)[24].

This granules are liberated after the recognition between target and T cells, the creation of pores unstabilize the cell, allowing proteases to reach the mitochondria and DNA.

Citotoxic T Cells also have an important role in the defense against virus and tumor growth through the secretion of IFN- γ and TNF-α [24].

B cells (CD19)

In contrast to citotoxic T cells, Bcells are in charge of humoral immunity, therefore the defense against the "outside" menace. B-cell antigen receptor is the surface immunoglobulin (or antibodys), an integral membrane protein with regions that allow them to recognize and bind

limitless antigens (specific molecules or patterns) in a specific way. Thousands of identical copies of antibodies can be found on the surface of B cell, composed by a common domain and a variable domain, the variable region allow the antigen-binding process through non- covalent forces. Therefore the 3D structure of the light and heavy chains that conform this domain play a mandatory role[25].

This antigen recognition triggers the activation of B-cells, due to the association of C-terminal portion of the antibodies with protein kinases[24]. This signaling process provokes a morphological change in B cells that multiply to become secretory factories of soluble antibodies.

Activated proliferating B cells that present more affinity for the antigen are selected for further replication and multiplication, providing a stronger response in long-lasting infections, with refined antibody specificity. When the menace is eliminated, depending on the antigen, the adjuvant and the infection route, distinct types of memory B cells can be gen-erated. Those clones of B cells with higher antibody specificity will be conserved as quiescent B cells, that will execute a quicker and stronger response in further exposures to the same antigen, as each infection will increase the amount of existing quiescent B cells[26].

Natural Killer (CD56)

NK cells constitute a third lymphoid line derived from a common, despite of that they do not present specific receptors for antigen recognition[27], either TCR or antibodies and not necessarily maturate on the thymus. NK cells have the highest cytotoxic capacity[28] and mediate the defense against intracellular pathogens and tumoral cells. Furthermore, activated NK cells are able to secrete matrix metalloproteinases (MMPs) for tissue remodeling, and also regulate of immune response. They are able to self-renew and proliferate and after that they return to a quiescent state [29].

Lymphoid Progenitor Cells (CD34+)

Also referred as Hematopoietic stem cells, lymphoid progenitor cells are pluripotent cells that can produce mature immune cells such as erythrocytes, leukocytes, platelets, and lymphocytes undergoing multiple divisions[30]. They are produced in the bone marrow, where they stay much of the time until they migrate into the blood or other tissues, differentiate into components of the blood or overcome apoptosis. They have an incredible grade of plasticity being able to transform even into epithelial cells.

Methodologies to identify VEGs

Different computational approaches and statistical models have been developed to identify VEGs from scRNA-Seq. Considering that in scRNA-seq count data, expression mean and variance are positively correlated different approaches have been taken in order to avoid VEGs to be a mere representation of those highly expressed genes[11].

scVEGs

scVEGs algorithm starts by assuming that scRNA-Seq follows a binomial distribution[7], where α represents biological dispersion and β represent the different sources of technical variation parameter, being therefore proportional to mean expression(μ) and constant through all the dataset. Considering that, it obtains α and β parameters regressing log10(CV) on the log10(mean), using robust local regression (locfit package), finally it implements nonlinear least-squares after subsampling the fitted data points to avoid overfitting of the regression[7] (Figure 5).

$$\sigma_i^2 = \frac{\mu_i}{1-p_i} = \mu_i + \frac{\mu_i^2}{r_i}, \longrightarrow \sigma^2 = \beta\mu + \alpha\mu^2,$$

Figure 4: Mathematical representation of variability distribution in scRNA-Seq according to the binomial distribution of variability. In scVEGs methodology α represents biological dispersion and β represent the different sources of technical variation parameter

Finally, the variability of the genes is determined by the difference between the modelled variability for the observed mean, and the observed variability. Finally this differences are assumed to follow a normal distribution, fitted using kernel density estimate (Figure 5). Given that, those genes with a pvalue <0.05 according to the obtained distribution are selected as VEGs.

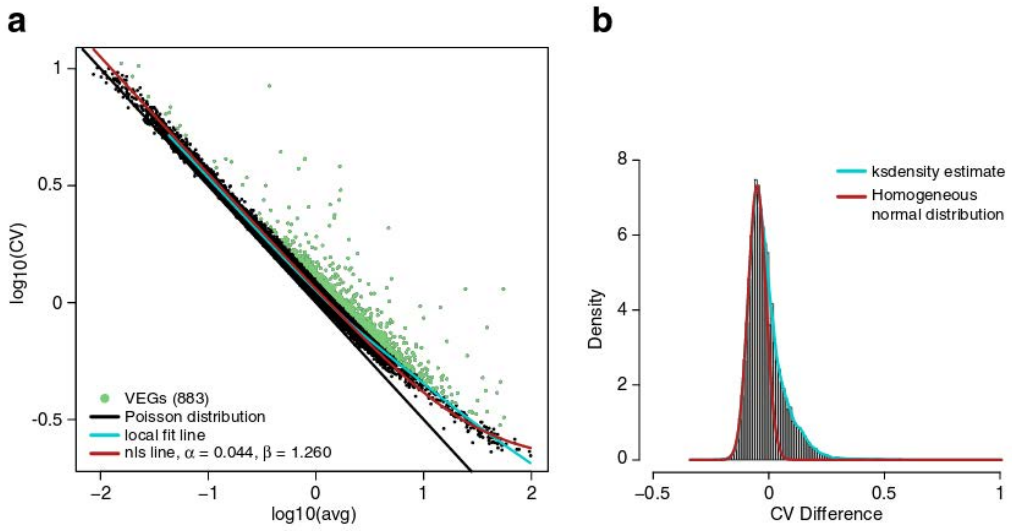

Figure 5 Representation of scVEGS procedure for expected variance calculation through regression adjustment on a log10CV vs log10mean plot and normal distribution adjustment (using kernel density) over the histogram of observed distances between VEGs (green dots) and expected variance (red regression line, nls)[7].

Seurat

Seurat[31] encloses three different methods for VEGS assessment. In typical SC Data analysis workflow, VEGs are used as features to cluster different cell population by mean expression. Therefore performing a classic differential expression analysis on average expression but focusing in those genes that present more variability within the sample.

VST (Seurat::VST): In a similar way to scVEGS, SEURAT::VST method Fit local polynomial regression LOESS to adjust a log(variance) log(mean) plot. LOESS perform a smooth regression

curve adjusted to the observed data, establishing the curve values as the expected variance for those genes with a given mean value. Finally, a standardized variance is calculated, as the ratio between the observed variance of each gene and the expected variance calculated by the model. This procedure was recently implemented in Seurat 3.0.2 release (June 2019).

Mean.Var.Plot (Seurat::MVP): calculate average expression and dispersion for each feature. Next, divides features into 20 bins based on their mean.exp , calculates z-scores for dispersion within each bin to determinate VEGs. Identify variable features while controlling relationship between variability and mean.

Dispersion (Seurat::DISP): This methodology calculates de Standard deviation of genes (SD) and select those n top features according to a given parameter (established on 2000 by default).

OBJECTIVES

In this project, we aim to characterize whether gene expression variability can characterize cells and tissues in both scRNA-Seq data (of clonal and more complex populations) and bulk RNA-Seq data of blood samples. Methods developed for the identification of VEGs from scRNA-Seq data will be applied to both Bulk and scRNA-Seq data.

MATERIAL AND METHODS

Datasets

In order to address the biological question we have put together the following datasets of immune cells obtained from human blood samples (Table 2). As a whole they are particularly interesting, as they have the particularity of being different pure cell populations selected by flow cytometry (Fluorescence-activated cell sorting, FACS) using the same membrane markers for the gating strategy in both Single cell RNA Seq and Bulk RNA Seq experiments. There is data from fully differenciated and pluripotent immune cells (as we have CD34 stem cells and fully developed lymphocytes (B & T Cells…NK). Also from cells with different degrees of relationship (for example, T Helper and T Citotoxic cells are closely related) and even we have cells from myeloid (CD14), and lymphoid lineage (CD19,CD4,CD8,CD56). Finally, all of the samples retain similarity for being part of the PBMCS cell class, we have considered that identifying cell subtypes from very differentiated cell types by VEGs could be deceiving, as we would observe relevant differences just by randomly sampling part of the expressed genes.

Finally, as the cherry on the top, we have an extra dataset for the whole PBMCs as a representation of a heterogeneous sample with the same expression profile.

Cell Type	Dataset Name	Samples (SC)	Genes/cel (SC)	Samples (Bulk)	Genes/Sample (Bulk)	Composition
PBMCS	PBMCS	2700	817	13	23462	Heterogeneous
Monocytes	CD14	2612	382	12	22493	Homogeneous
B Cells	CD19	10085	478	20	23354	Homogeneous
T Helper Cells	CD4	11213	546	30	23624	Homogeneous
Natural Killers	CD56	8385	710	4	27420	Homogeneous
T Citotoxic Cells	CD8	10209	573	16	22626	Homogeneous
Progenitor cells	CD 34	9232	1274	4	27420	Homogeneous

Table 2: Characteristics of the different datasets (from SC-RNA-Seq and Bulk RNA-Seq data) that were employed in the present manuscript. SC RNA-Seq data was obtained from 10x Genomic repository (Cell Rangel 1.0.1) https://support.10xgenomics.com/single-cell-gene-expression/datasets and Bulk RNA-Seq data from GEO accession number GSE107011.

Bioinformatic procedures

R session info:

```
R version 3.4.4 (2018-03-15).
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows >= 8 x64 (build 9200)
```

The code for implementation of all the analysis (except from IPA for copyright issues) can be found on the R Markdown provided as Supplemental Material. The enclosed zip file also provides the complete session info, original datasets and necessary results to perform all the procedures in a standalone way.

Supplemental Material can be found in the following url: https://drive.google.com/open?id=1danah76vp_aAFcm0eWH0-JE3Pud8xD-O

SC data processing

SC Data was downloaded from 10x genomics public repository. https://support.10xgenomics.com/single-cell-gene-expression/datasets/ (Cell Ranger 1.0.1) and processed with Seurat 3.0.2 package (release Jun 2019). Datasets were filtered for those cells presenting over 2500 or under 200 counts and with a mithocondrial counts percentage above 5% according to Seurat authors standard; after data was normalized using normalization method "Log-Normalized" (a Seurat function that normalizes the expression values of each cell by the total expression, and multiplies this by a scale factor (10,000 by default), after that resulting values are log-transformed (natural logarithm)), and processed according with the R Markdown provided in Supplementary Files.

Bulk Data processing

Bulk RNA-Seq data from TPM Normalized data was downloaded from GEO accession number GSE107011 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107011 (Supplementary File) and grouped according to authors nomenclature for each kind of cell [32].

Data was log normalized (performing the nature logarithm of 1+(expression matrix)) and genes expressed in less than 70% of the samples for each of the cell datasets were filtered out.

Genes were annotated as hgnc gene symbols using ensemble database and biomaRt package as described in Supplementary Material

Seurat VEG procedures

For Dispersion procedure, the default parameter of n = 2000 was used.

For Bulk processing the filtered and log normalized TPM expression data was loaded as a Seurat object avoiding any kind of preprocessing, raw.data slot was used as input for the VST FindVariables.

scVEGS

scVEGS was run using the recommended default parameters (pval=0.1, pFlag= 1).

For Bulk processing, the scVEG original function was modified in order to avoid the TPM transformation.

Gene Set Enrichment analysis

Functional annotation was performed using Ingenuity Pathway Analysis (QIAGEN Bioinformatics) Summer 2019 Release. VEG lists were contrasted to the whole Ingenuity knowledge database without using any value as Z-score substitute.

Statistical Analysis

Study of enriched molecular types and cell locations was performed using Exact Fisher test of each cell line VEGs list with the matching complete dataset (variably and non-variably expressed genes) without filtering from the lists those genes with unknown function/location (labeled as “others”). P-value post-hoc correction was performed using FDR procedure, in this case we did not include the comparison of genes labeled as “others” in the correction as they had no biological interest.

RESULTS

Comparing methodologies

Single cell

As reported in the literature, the outcome of VEG determination algorithms (including those tested in this manuscript) are heterogeneous and depend on the datasets characteristics [12] (Table 3). In our experiment, all the algorithms determined a higher amount of VEGS in the most heterogeneous dataset, the PBMCS, and shared a significant amount of genes between them (Figure 6) (Of course this was not observed in Seurat::DISP, which always returns a fixed amount of features determined by its "n" parameter).

SCData	SEURAT::VST	scVEGS	SEURAT::MVP	Seurat::DISP
CD14	470	54	520	2000
CD19	106	91	472	2000
CD34	264	431	400	2000
CD4	160	158	288	2000
CD56	99	191	413	2000
CD8	88	211	356	2000
PBMCS	1165	2090	995	2000

Table 3: Obtained VEGs in SC-RNA-Seq datasets with each of the tested methodologies. Note that Seurat::DISP function requires an output parameter that determines the number of obtained features, 2000 by default.

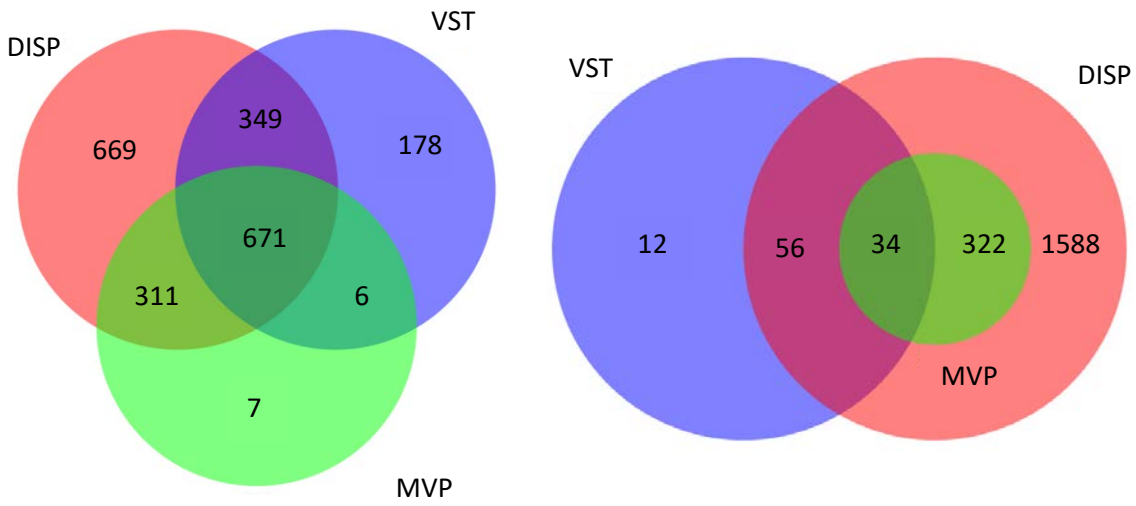

Figure 6: Venn Diagrams of VEGs lists obtained on SC-RNA-Seq with Seurat methodologies (SEURAT::VST (Blue), SEURAT::MVP (Green) and Seurat::DISP (Red)) on PBMCS (Right) and CD8 (left). Note that circle sizes are not proportional by any mean.

In all of the situations, SEURAT::MVP VEG are almost contained in the set obtained by the dispersion method, this would suggest that dispersion and SEURAT::MVP method are performing similarly, with the dispersion method being less strict in the selection of the most variable genes and therefore more prone to type I error. In addition, and because the dispersion method has no threshold for feature selection, it always returns 2000 features (as the default Seurat parameter).

Regarding scVEGS and SEURAT::VST (being two methods that model the expected dispersion in order to determine which genes present a higher variability, and getting rid of the impact of average expression) and worked smoothly (Figure 7). Both delivered gene lists that shared a high complementarity in PBMCS datasets, being more divergent in more homogenous samples (Figure 7).

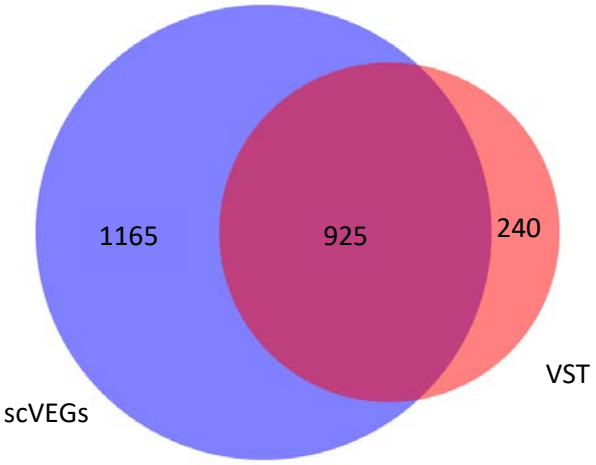

Figure 7: Venn Diagrams of VEGs lists obtained on SC-RNA-Seq with Seurat SEURAT::VST and scVEGs methodologies (scVEGS (Blue) and SEURAT::VST (Red)) on PBMCS.

Bulk

In order to be able to analyze Bulk-RNA-Seq data with ScVEGs package, a modified scVEG function was applied in order to avoid a double TPM normalization, allowing the input of Bulk RNA-Seq data matrix performed correctly. With this modification the method performed correctly but it didn't detect any VEGs in those datasets, probably as a consequence of the highest average expression values obtained on the Bulk data methodology and the smaller

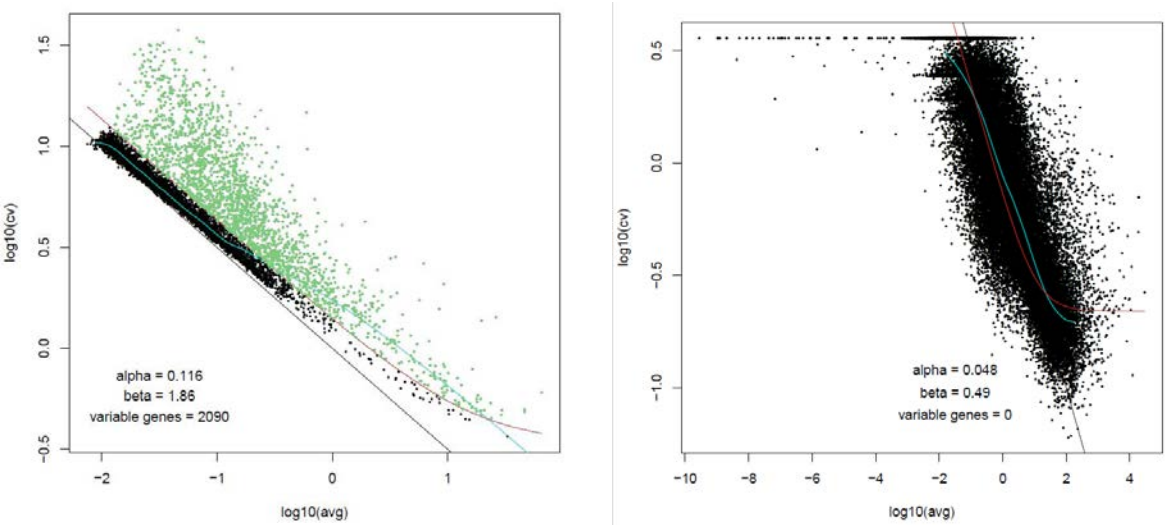

Figure 8: Output results of applying scVEGs methodology in SC-RNA-Seq (right) and Bulk-RNA-Seq (left) PBMCs datasets. Green dots represent features identified as VEGs. Scales are not standardized.

dispersion observed (measured as CV) (Figure 8). This output did not changed regardless of the homogeneity or sample size of the analyzed dataset.

Seurat methods:

Out of the three variability measurement methods implemented in Seurat, SEURAT::VST was the only one that performed correctly when a bulk TPM log-normalized matrix was given as input. Using 1.5 threshold for the adjusted variance for feature selection as in the case of single cell data returned very large lists of variables genes. Given that, we decided to apply a more restrictive cutoff point of 3.0 (meaning that the genes would have to showcase 3 times more variance than the variance estimated for a gene with the same expression).

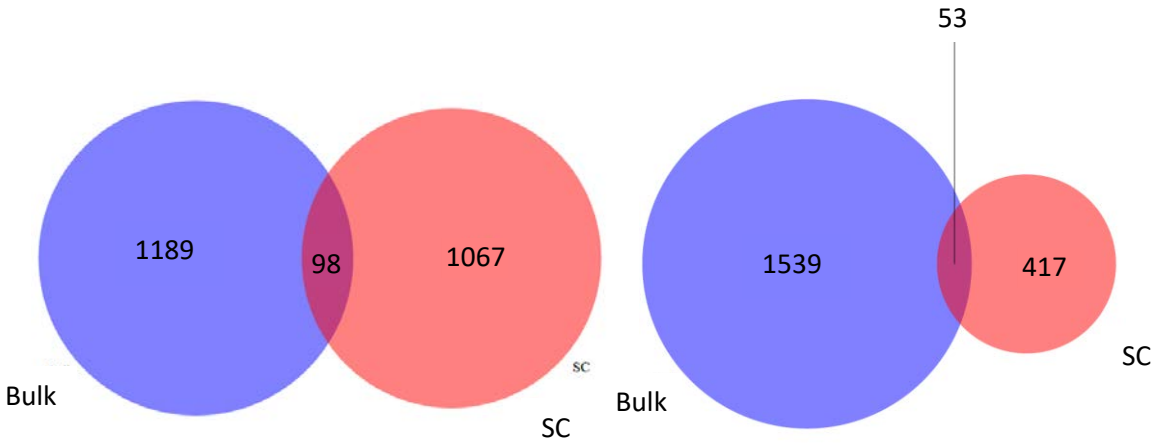

Figure 9: Venn Diagrams of VEGs lists obtained on Bulk (Blue) vs SC-RNA-Seq (Red) using Seurat VST methology on PBMCS (Right) and CD14 (left)

When compared with the SC results for the same cell type it can be observed that even in the most heterogeneous samples (PBMCS) the obtained lists share a very low concordance with their SC counterparts (Figure 9). This is especially relevant considering that the Bulk-derivated gene lists contain more than 1.000 features for every cell type (Table 4). In this datasets, the total number of samples per group (n=4 in CD34 and CD 56) is the main factor affecting the amount of obtained VEGs.

cell_line	VEGS_Bulk	VEGS_SC
CD14	1592	470
CD19	1029	106
CD34	2586	264
CD4	1517	160
CD56	2227	99
CD8	1743	88
PBMCS	1287	1165

Table 4 Obtained VEGs in SC-RNA-Seq and Bulk-RNA-Seq datasets using SEURAT::VST methods.

Taking into consideration the observed divergence between the different methods (as previously reported[12], we decided to choose a methodology that could be used in both datasets in order to reduce the methodological bias. Particularly, if we expect to compare SC vs Bulk data is mandatory to choose a methodology that allow us to correct the value of dispersion in regards of the mean expression, due to the inherent differences between bulk RNA sequencing and Single Cells RNa Sequencing (with an increased amount of unexpressed genes and lower intensities) . Considering this, we decided to choose SEURAT::VST as the reference methodology.

Comparing variability profiles

We have contrasted the standardized variability obtained with Seurat::VST of paired datasets (SC vs Bulk) to compare the variability profile obtained with both techniques.

Interestingly we have also found that despite of scRNA-seq, where expression mean and variance are positively correlated, in Bulk RNA Seq they are inversely correlated.

Comparing only the common genes in Bulk and Single cell datasets for each cell type (Figure 10, 11) we can have a more accurate comparison of the actual variability distribution differences between techniques, considering that the total genes per cell in SC is more limited than in bulk sequencing(Table 2). We have also plotted the profiles using the same x and y scale in order to get a better picture of the profiles to be compared (outliers removed). Outliers in Bulk variability distribution are a constant feature that do not get represented in this kind of histograms. (Figure 10).

In this plot we can observe that PBMCS variability present a wider distribution, more similar to a log normal plot, with values way more distributed along the x axis in a more continuous way while the SC plot (negative binomial) have a distribution more centered around the 1.

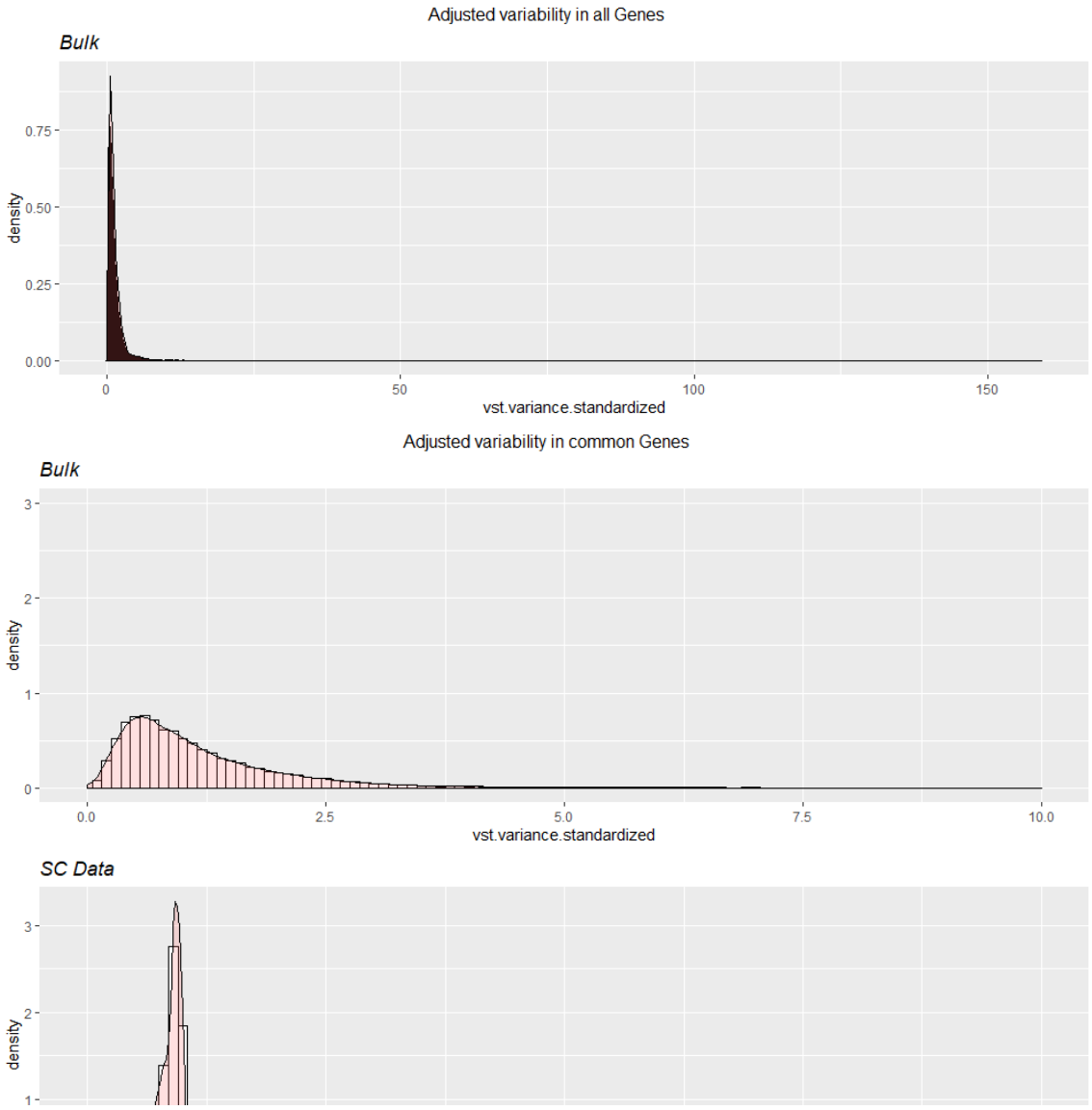

Figure 10: Variability histograms of common genes between Bulk (top) and SC-RNA-Seq (bottom) complete dataset of PBMCS (not limited to VEGs). Standardized variability is calculated by VST method dividing observed variability between expected variability for a specific gene regarding their level expressions. Please note that in x and y axis have been homogenized for comparison meanings. At the very top it can be observed the original PBMCs graphic, shrinked by the presence of outliers in high values and shorter y axis.

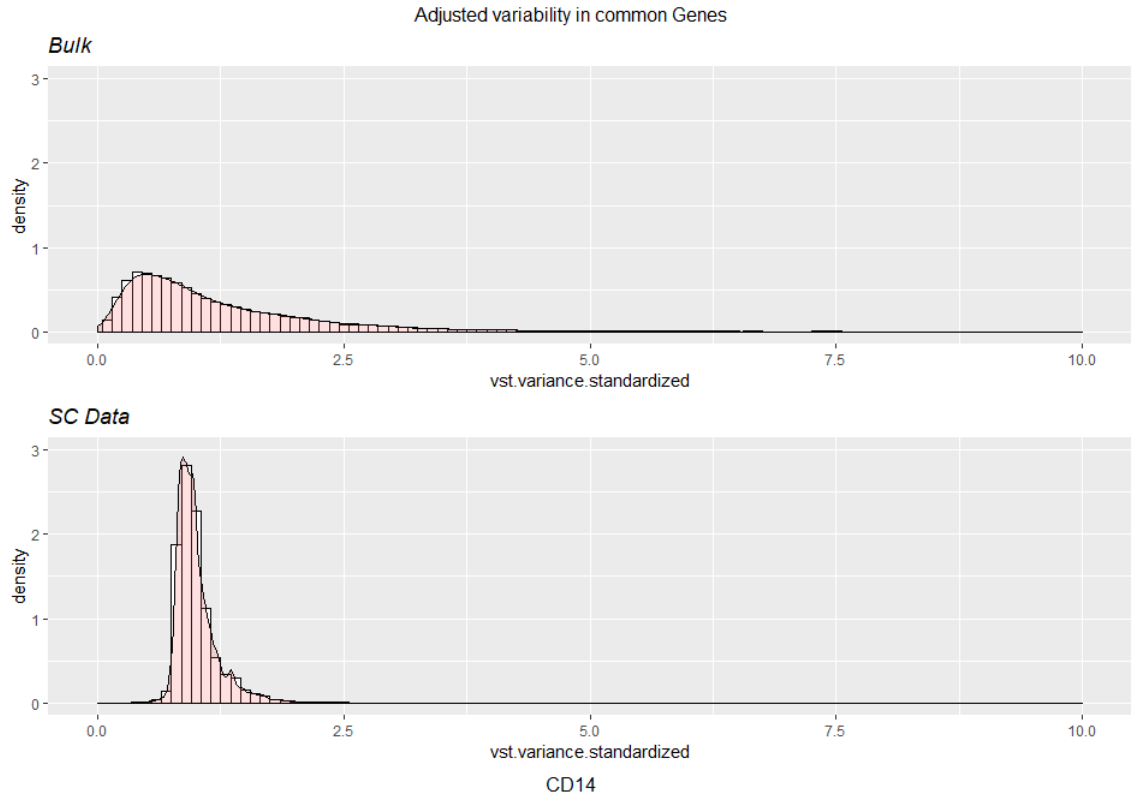

Figure 11: Variability histograms of common genes between Bulk (top) and SC-RNA-Seq (bottom) complete dataset of CD14 (Monocyte) cells (not limited to VEGs). Standardized variability is calculated by Seurat::VST method dividing observed variability between expected variability for a specific gene regarding their level expressions. Please note that in x and y axis have been homogenized for comparison meanings, omitting Bulk.RNA-Seq standardized variability outliers

A common feature that can be observed in the bulk variance study using SEURAT::VST methodology is the existence of many variability outliers, resulting in something close to a log normal distribution, with no-negative values and a long tail. The appearance of outliers would be accentuated by the discrete amount of samples within each cell type in bulk datasets (compared to SC). As it can be observed in the Table 5 the outliers are generated because of the low average variability observed in Bulk data (calculated with the average 30.000 genes of each sample), and the innacurate variability measures obtained in small datasets with a number of samples that ranges from 4 to 30 (Table 2) .

	ENSGSplit	vst.mean	vst.variance	vst.variance.expect~	vst.variance.standar~	hgnc_symbol
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	ENSG00000257~	6.02	10.7	0.0490	159.	NA
2	ENSG00000276~	3.93	5.48	0.0509	84.2	NA
3	ENSG00000067~	3.57	5.85	0.0572	79.4	DDX3Y
4	ENSG00000237~	4.60	4.79	0.0442	74.6	NA
5	ENSG00000036~	4.26	3.43	0.0467	55.9	MYOM2
6	ENSG00000211~	4.88	2.36	0.0437	53.1	TRBJ2-7
7	ENSG00000090~	4.28	2.46	0.0465	40.2	RGS1
8	ENSG00000012~	3.02	3.33	0.0761	39.3	KDM5D
9	ENSG00000254~	3.54	2.45	0.0579	39.1	SIGLEC14
10	ENSG00000211~	4.13	2.15	0.0481	37.8	IGHJ1

Table 5: Variability outliers present in Bulk-RNA-Seq PBMC dataset processed by SEURAT::VST methodology

In addition, Bulk variability presents a wider distribution in the low values, achieving smaller densities than the single cell counterpart.

In the other hand the SC, data distribution of variability is defined as a negative binomial, as described in the literature[9] without remarkable outliers and a higher density of values around 1.

Comparing the absolute variability deviation from SC to Bulk for each the gene in paired datasets SC-Bulk (Figure 12, 13) we can observe that the difference is centered on a value close to 1 and distributed in a normal-like way in PBMC. In more homogenous datasets, the distribution of difference values embiggens, centered in values closer to 1. In this scenario, a difference of 1 in the adjusted variability would mean that the Bulk observed variability will double the expected one for it means in comparison with the SC screening for the same cell type).

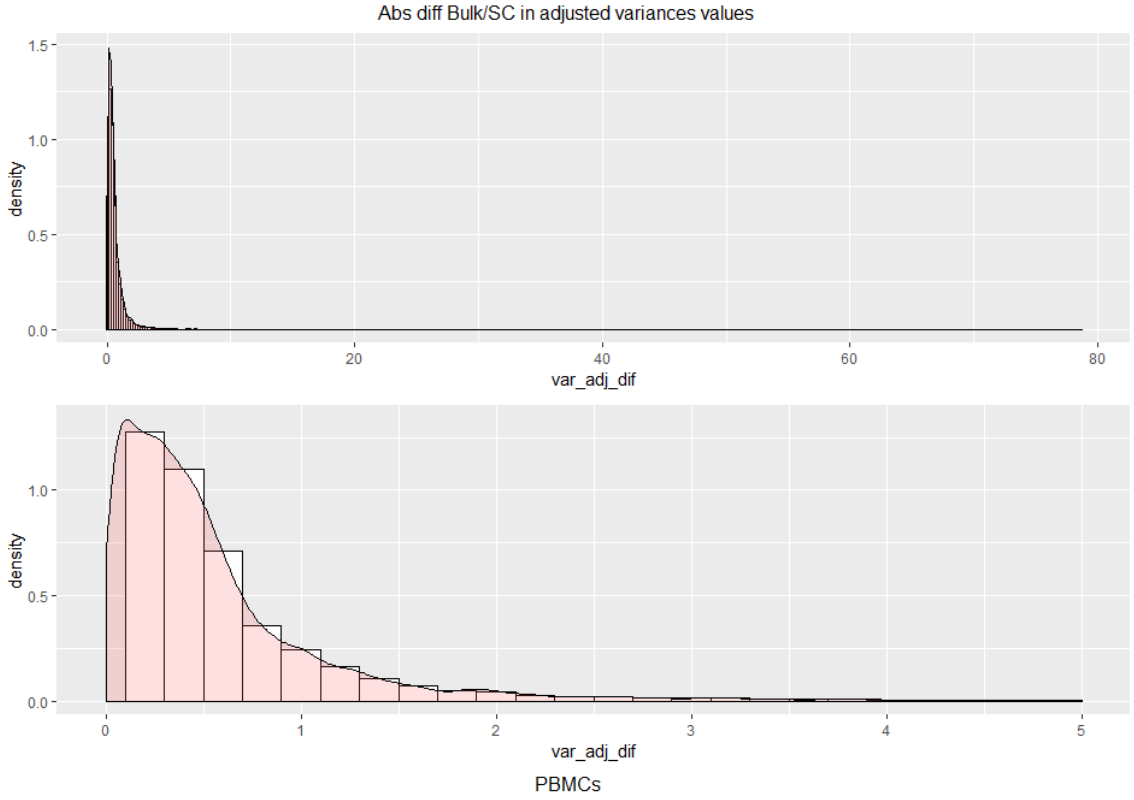

Figure 12: Histogram representing the distribution of the absolute difference between adjusted variance obtained in SC-RNA-Seq data vs Bulk-RNA-Seq data in the same cell type (PBMCs) for each of the genes expressed in both datasets. Standardized variability is calculated by SEURAT::VST method dividing observed variability between expected variability for a specific gene regarding their level expressions.

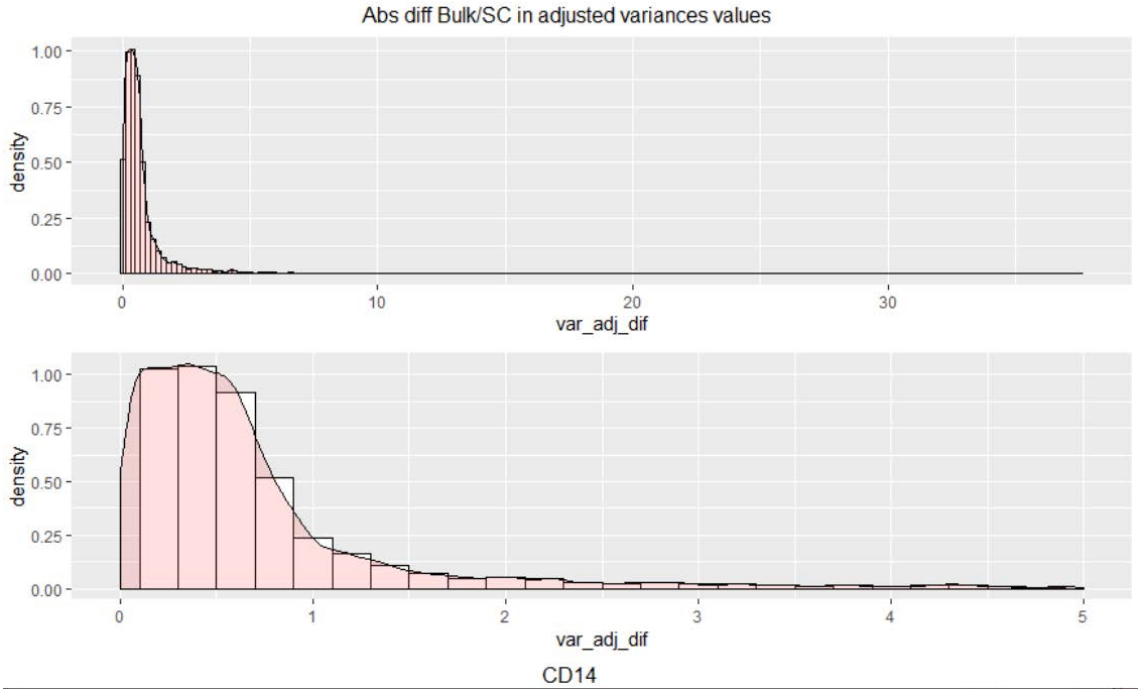

Figure 13: Histogram representing the distribution of the absolute difference between adjusted variance obtained in SC-RNA-Seq data vs Bulk-RNA-Seq data in the same cell type (CD14) for each of the genes expressed in both datasets. Standardized variability is calculated by SEURAT::VST method dividing observed variability between expected variability for a specific gene regarding their level expressions

Biological characterization through Variability in SC

Gene Set Enrichment Analysis

Functional enrichment analysis were performed using the lists of VEGs for each cell dataset. IPA (Ingeniuty Pathway Analysis) software was used and each dataset was contrasted to the whole IPA knowledge base database, in order to determine if the VEGS could characterize the biological activity of the cells. Moreover we were interested in determine whether the VEG profiles could allow us to characterize each of the different cell types. (Figure 14). In fact, we can observe that genes that present a high variability in their expressed profiles are related with biological processes involved or related with immune response.

Figure 14: Dot plot representing the top 4 Enriched pathways in VEGs (identified by GSEA analysis of SEURAT::VST VEGs gene lists in SC-RNA-Seq data) for each cell type in SC-RNA-Seq data analyzed with SEURAT::VST method. Dot size is proportional to the number of VEGs identified for each pathway and heat colour represents BH correction p-value (representing red the smaller pvalues). Grey dots represent pathways that were identified on the cell lines VEGs but not in a significantly enriched after BH correction.

Despite of having the longest list of VEGs (and considering that we performed a simple over-representation analysis without associated Z scores). PBMCs show smaller log values than those more homogenous samples as CD34, CD4 and CD14.

In contrast, no enriched pathways were found for CD19 after Bonferroni-Hochberg multiple testing correction is applied; this could be a direct consequence of having the smallest VEGs list among all the cell lines. This seems like a feature of the cell line itself, as the technical parameters of this dataset do not differ from the other in terms of total cells (10.000) or average reads per cell (25.000). CD56 (NKs) also present a slightly lower variability than the other lymphoid cells.

CD8 (Citotoxic T cells) presents a profile of intermediate variation, similar to PBMCs. This can also be observed if we compare those pathways that are significant in more than 4 cell lines (Figure 15).

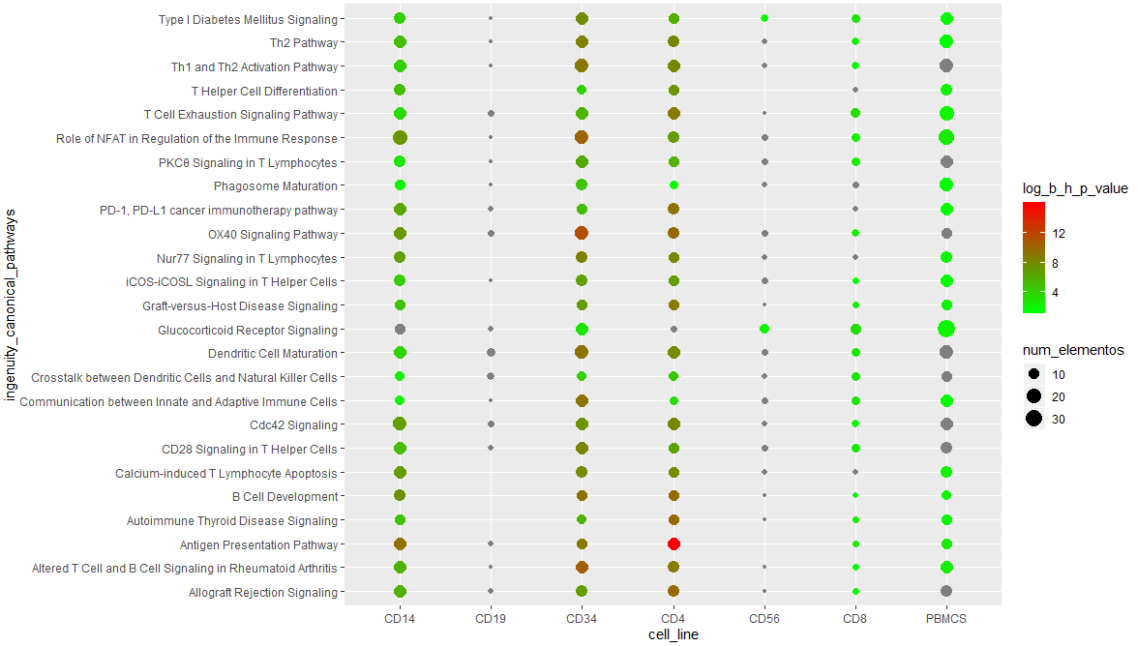

Figure 15: Dot plot representing pathways that were significantly enriched in VEGs lists in more than 4 of the 7 cell lines in SC-RNA-Seq data analyzed with SEURAT::VST method. Dot size is proportional to the number of VEGs identified for each pathway and heat colour represents BH correction p-value (representing red the smaller pvalues). Grey dots represent pathways that were identified on the cell lines VEGs but not in a significantly enriched after BH correction.

Observing those pathways that are significantly enriched exclusively in one of the different cell types, but only considering the top 30 pathways for each cell. (Figure 16). We can observe how CD14 and PBMCs have a more divergent variability profile, for being either a cell type from a different lineage (or a heterogeneous cell sample (where variability arises). Due to the shared lymphoid lineage of all the other cells, (or the actual non-enrichment as in the case of CD19) there are no many enriched unique functions in only one cell type.

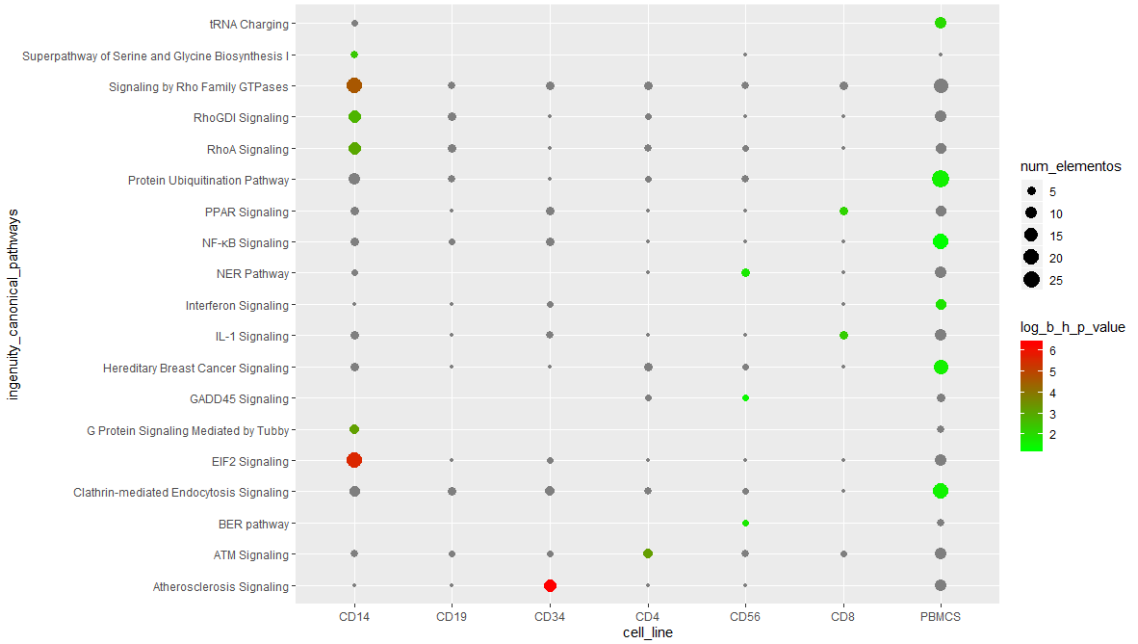

Figure 16: Dot plot representing pathways that were significantly enriched in VEGs lists exclusively in 1 of the 7 cell lines and also between the top 30 results for each cell line in SC-RNA-Seq data analyzed with SEURAT::VST method. Dot size is proportional to the number of VEGs identified for each pathway and heat colour represents BH correction

p-value (representing red the smaller pvalues). Grey dots represent pathways that were identified on the cell lines VEGs but not in a significantly enriched after BH correction.

If we take a closer to the top 4 enriched pathways we can find out very interesting results in terms of biological characterization (despite of the fact that a lot of features and pathways are shared through several cell types due to their similar lineage, making this analyze more challenging). For example, Granzyme signaling is enriched in NK and T cytotoxic cells, granzyme is a serines protease that are secreted by NK and CD8 cytotoxic granules as a mechanism of defense against tumoral and viral infected cells[33]. [34]

Il17A signaling, a interleukin involved in T cell differentiation and regulatory mechanisms (Figure 3) is exclusively enriched in CD4 and CD8 (T cells). Also, antigen presentation pathways are enriched in T helper, T cytotoxic and monocytes, being those cell types involved in the antigen presentation procedure (Figure 3).

Monocytes presents several enriched pathways involved in Rho GTP-ase pathways, this molecules are involved in monocyte motility response through tissues [35]. Being monocytes the most distinctive of those homogeneous and differentiated cell types, they are also easier to characterize, having several Rho-related pathways exclusively enriched in this cell type (as there are no other myeloid cells in our dataset) (Figure16). Similarly, monocytes and NK cells presents enriched pathways in remodeling epithelial adherens junctions, as they must infiltrate tissues as part of their biological functions [15].

Biological features enrichment

We performed a secondary analysis of biological significance across the immune cell types. We annotated the obtained VEGs in terms of both molecule type and cell location using Ingeniuty pathways database in order to determine if those biological features were enriched in VEGs in contrast to the whole whole geneset in any of the cell types.

Results can be consulted in Table 6. As expected, CD 34, progenitor stem cells, showcase the higher enrichment in terms of variable genes in both cell location and biological functions, being undifferentiated cells with high plasticity.

cell_line	location	FDR
CD14	Plasma Membrane	2.443061e-04
CD34	Extracellular Space	4.212309e-22
CD34	Cytoplasm	1.204404e-02
CD34	Plasma Membrane	1.339366e-06
CD34	Nucleus	1.401476e-05
CD4	Cytoplasm	4.665996e-02
CD4	Plasma Membrane	1.082282e-04
CD4	Extracellular Space	7.035950e-05
CD56	Extracellular Space	1.339366e-06
CD8	Cytoplasm	3.702248e-02
CD8	Plasma Membrane	2.205372e-02
CD8	Extracellular Space	7.035950e-05
PBMCS	Cytoplasm	8.230088e-03
PBMCS	Extracellular Space	1.834977e-02

cell_line	type_s	FDR
CD14	transmembrane receptor	1.069890e-09
CD14	kinase	3.907363e-03
CD34	enzyme	1.102541e-03
CD34	peptidase	8.226743e-04
CD34	cytokine	2.651532e-11
CD34	transmembrane receptor	2.651532e-11
CD34	kinase	2.067659e-02
CD4	cytokine	8.953838e-03
CD4	transmembrane receptor	2.862163e-09
CD56	cytokine	1.442411e-04
CD8	cytokine	8.226743e-04
CD8	transmembrane receptor	5.146542e-03
PBMCS	transmembrane receptor	1.496498e-02
PBMCS	cytokine	8.226743e-04

Table 6: Results of biological features enrichment analysis of VEGs for celular location (right) and molecular function (left) after FDR correction. VEGs were obtained through VST methodology in SC-RNA-Seq Data.

CD14 (Monocytes), the only cell of myeloid lineage in the datasets have a representative pattern of variability in both cell location and molecular type, whereas cells from lymphoid lineage (CD 8, CD4 and CD56) also shares a distinctive and common pattern that differs from that on the myeloid cells.. In terms of molecule function, Cytokynes are enriched in lymphoid cells, something biologically representative of this immune lineage that secrete cytokines to mediate their activity and regulation on other immune cells, in the other hand kynases are over-represented in the only myeloid related cell, the CD14.

Regarding CD 34 (Table 6) they present enrichment in most of the locations, as it was observed in the molecular type data. Regarding the lymphoid lineage, they share the over-representation of extracellular components.

In both cases, CD8 and CD4 (cytotoxic T lymphocytes and T helper lymphocyte), those most tightly related cells share a yet another common pattern of enriched features in their VEGs), regarding both functions and locations with a higher representation of cytokines, transmembrane receptors and extracellular components. This characteristic is also observed in the PBMCS. In the case of PBMCs, being a heterogeneous cell type, part of their VEGs are a reflection of that inner heterogeneity, but as they are composed of mostly T cells (60% of total PBMCS (Table 1). Their functional enrichment within those VEGs summarize the most relevant features of T Cells. Finally, B cells (CD19) show no enrichment on either types of molecule or cell location, as observed in the pathway enrichment analysis, this could be a consequence of the quiescent status of B in the absence of external antigens or just a characteristic of the obtained sample. Further analysis indifferent conditions should be done in order to address this.

Biological characterization through Variability in Bulk

In bulk data the enriched pathway analysis returned a larger amount different pathways, as the VEGs list were more extense (Table 4), most of those pathways are related with the immune system. But in contrast with the single cell data, they show almost no difference between cell

types, showcasing similar patterns and pvalues. Despite of that, some interesting results can be cherry-picked in the top5 most enriched pathways for each cell type (Figure 17) but doesn't seem like we are being able to characterize cell types and we are just observing immune cells expression patterns.

Figure 17: Dot plot representing the top 5 Enriched pathways in VEGs (identified by GSEA analysis of SEURAT::VST VEGs gene lists in SC-RNA-Seq data) for each cell type in Bulk-RNA-Seq data analyzed with SEURAT::VST method. Please note that in this case top 5 was considered due to the homogeneity of the results.

Finally, the analysis of the top 30 pathways that are enriched in a single cell type and not in the rest of the cells (in order to observe unique patterns of variability) returned one single pathway in one cell type without any biological relevance. Increasing that value to 100 (because the higher amount of VEGS reported a large amount of enriched pathways) allowed us to obtain some results, but yet again, they are not very promising. (Figure 18).

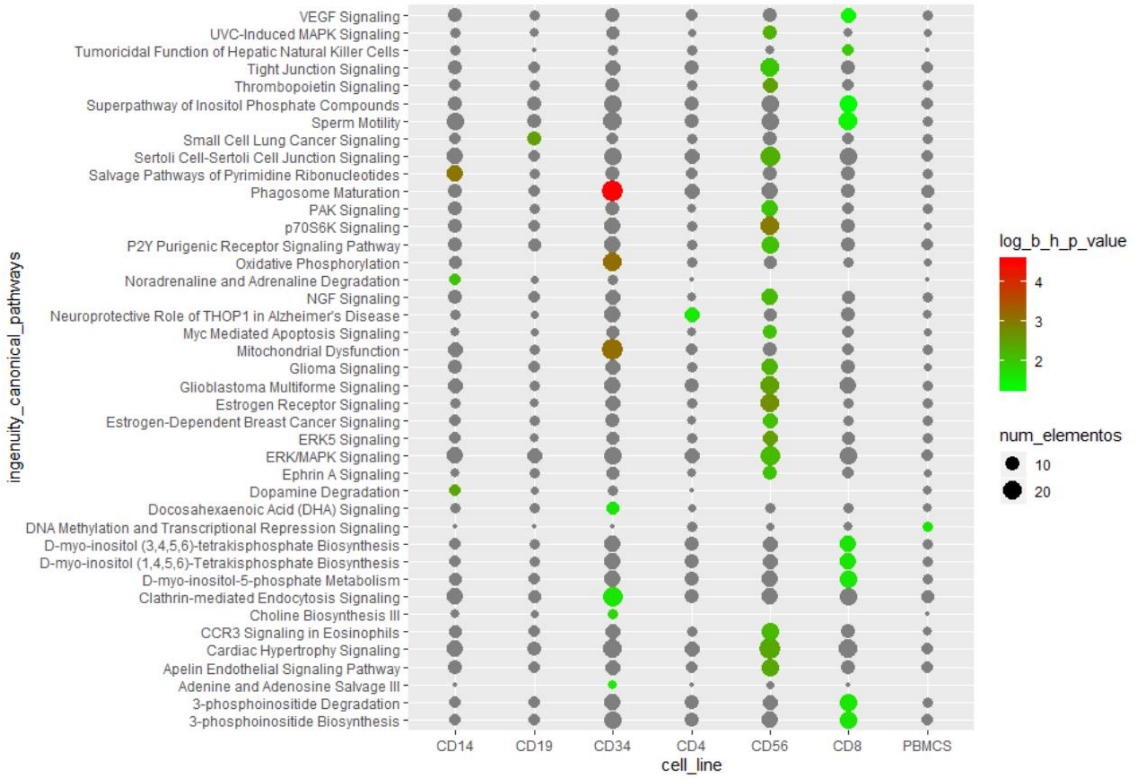

Figure 18: Dot plot representing pathways that were significantly enriched in VEGs lists exclusively in 1 of the 7 cell lines and also between the top 100 results for each cell line in Bulk-RNA-Seq data analyzed with SEURAT::VST method. Please note that in this case the number of results has to be set from 30 to 100 in order to obtain results due to the homogeneity of the results.

Regarding the functional enrichment analysis, similar outcome was obtained, with almost every function and locations enriched for all of the different cell lines, without any kind of difference regardless of function, differentiation status or relationships. Only a similar over-representation of locations was found on T Cells (CD4 and CD8). In the case of molecular function comparison, all the cell lines presented the same pattern with almost every function over-represented. (Table 7).

Cell Line	Location	FDR
CD14	Plasma Membrane	5.60E-14
CD14	Cytoplasm	1.11E+01
CD14	Nucleus	1.19E-15
CD14	Extracellular Space	3.62E+02
CD19	Plasma Membrane	1.72E-03
CD19	Cytoplasm	2.09E-02
CD19	Nucleus	2.07E-06
CD19	Extracellular Space	1.20E+01
CD34	Extracellular Space	2.62E+02
CD34	Nucleus	1.32E-32
CD34	Cytoplasm	2.34E+00
CD4	Plasma Membrane	6.47E-02
CD4	Cytoplasm	4.30E-05
CD4	Nucleus	1.19E-18
CD56	Cytoplasm	1.17E-08
CD56	Nucleus	4.22E-11
CD8	Plasma Membrane	3.35E-08
CD8	Cytoplasm	8.76E+01
CD8	Nucleus	1.14E-19
PBMCS	Cytoplasm	2.96E-17
PBMCS	Extracellular Space	4.47E-03
PBMCS	Nucleus	8.09E-14

Table 7: Result of biological features enrichment analysis of VEGs for celular location after FDR correction. VEGs were obtained through SEURAT::VST methodology in Bulk-RNA-Seq Data. Note that molecular function results are not being displayed because of the length of the table (54 entries) and no biological relevance.

DISCUSSION

Considering all the results presented in the current manuscript we can confirm that variability profiling allow us to characterize the biology of the sample in single cell data.

Methodologies

As previously reported[12] methodologies for VEG determination in SC sequencing have discrepancies in their result, and methods that minimize the effect of the direct correlation between mean and dispersion are necessary to avoid picking only those highly expressed genes. SEURAT::VST have proved to provide meaningful results when processing single cell datasets, and being to process Bulk RNA-Seq datasets.

Variability comparison

Regarding the variability profile of bulk and single cell data, instead of the double binomial distribution presented by SC dispersion data as reported in the literature[1] the bulk dispersion data presents a log normal distribution. Studying the difference between the adjusted variability we can address that there is a technical bias in associated to Bulk RNA-Seq data. This effect gets more accentuated in homogeneous cell cultures.

SEURAT::VST SC

The gene set enrichment analysis demonstrated that identified VEG genes are involved in different immune related processes across the different cell types. Those pathways present cell type specificity, with enriched features that reflect the biological function of the different cell types. The high similarity between datasets proves that VEGs profiling analysis are cell type specific in the case of single cell RNA-Seq.

VEG analysis suggest that those as genes prone to be adapted through external stimulus (for example as an micro-environmental response or a differentiation procedure)[2]. All of the studied features (pathways, gene expression location and molecular type) presented results that were coherent in terms of identifying related cell types (CD4 & CD8 similarities), different lineages (lymphoid vs myeloid patterns), stages of differentiation and heterogeneity (PBMCS) vs homogeneity of datasets. In the case of CD39, progenitor cells, they presented a more diverse variability pattern in terms of functions and locations, this would reflect not only the different stages of differentiation in which this cells can be found, but also the high adaptability and plasticity of this precursor cells.

Regarding PBMC dataset, because of the heterogeneity of the sample[15], the total amount of VEGs will be always superior than in homogeneous samples. In addition, considering the percentage of each immune cell type in PBMCs (Table 1), the functional enrichment of this cellular type is highly influenced by the lymphoid T cells (CD4 & CD8).

With all the provided data, we can conclude that variability measured by SEURAT::VST procedure allow us to characterize the biology of immune cells, delivering meaningful results that are not just associated with technical issues but biological procedures.

SEURAT::VST Bulk

When analyzing Bulk RNA-Seq data, the lack of difference between cell lines, heterogeneous and homogeneous datasets or related and unrelated cell lines, and the low concordance between Bulk and SC results points out that SEURAT::VST is not providing an accurate determination of VEGS in bulk data.

Despite of the good results obtained in SC to characterize the different cell types. The results obtained, unveils something that preliminary data analysis already pointed out, the SEURAT::VST procedure is performing a weak and selection on VEGS , with an unusual Type I error, in a similar way as the dispersion method would perform in single cell datasets. Therefore, ranking those genes that present high variability and being unable to control the effect of the mean expression. In the future, we will need to obtain specific methodologies that allow us to obtain the most variable genes in Bulk datasets while being able to correct the effects of the mean. There have been some simplistic approaches to the date[36] that function similarly to SEURAT::MVP procedure, creating bins of each average expression and finding the top variable genes within each bin, but more accurate methods must be developed, if we expect to obtain meaningful results.

Finally, the low amount of samples that are obtained in Bulk RNA seq experiments are still the major constraint in the labor of assessing VEGS, a statistical approach that requires a larger sample size than mean gene expression studies.

Future approaches

Considering the results presented in this manuscript were limited by the time available for the present realization (300h) and the diverse methodologies that were implemented in different datasets, there are many challenges that should be tackled in the future in order to keep on the present research line.

Develop a decision tree algorithm (supervised classifier) with the top 5 VEGs features for each cell type to classify scRNA-Seq dat. We started this procedure for the current manuscript, but having only 1 variation measurement for each cell type (despite of having 3.000 cells we get a single standardized variance result for each cell line). In order to approach this problem we would perform bootstrap variability calculation with random subsets of the original dataset in order to get over at least 100 variability measurements per cell line, and then perform the classifying algorithm.

Validate the reported SC results with SEURAT::VST methodology in an external dataset. In case of not being possible to find a similar dataset, perform a new study using a test/validation approach within the existing SC RNA-Seq datasets.

Use overlapping features between different methods (SEURAT::VST/scVEGS/SEURAT::MVP) as VEGs lists and compare the obtained results with the already reported results.

Obtaining a bigger dataset of Bulk-RNA-Seq (PBMCs and cell lines), to address the factor of sample size in a more objective way.

Finding a specific Bulk methodology to determine VEGs:

> Able to perform within the same condition (and not as a differential comparison between two conditions as MDSeq).

Not relying on modelling expected variability, to avoid the issues that arised as a consequence of low variability in Bulk RNA-Seq data.

Avoiding simplistic approaches that would lead to poor results for ignoring the mean bias in variability.

BIBLIOGRAPHY

[1] S. Roberfroid, J. Vanderleyden, and H. Steenackers, “Gene expression variability in clonal populations: Causes and consequences,” *Crit. Rev. Microbiol.*, vol. 42, no. 6, pp. 969–984, 2016.

[2] S. Ecker *et al.*, “Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types,” *Genome Biol.*, vol. 18, no. 1, pp. 1–17, 2017.

[3] R. Bueno and J. C. Mar, “Changes in gene expression variability reveal a stable synthetic lethal interaction network in BRCA2−ovarian cancers,” *Methods*, vol. 131, pp. 74–82, 2017.

[4] T. Hagai *et al.*, “Gene expression variability across cells and species shapes innate immunity.,” *Nature*, vol. 563, no. 7730, pp. 197–202, 2018.

[5] Y. Hasegawa, D. Taylor, D. A. Ovchinnikov, E. J. Wolvetang, L. de Torrenté, and J. C. Mar, “Variability of Gene Expression Identifies Transcriptional Regulators of Early Human Embryonic Development,” *PLoS Genet.*, vol. 11, no. 8, 2015.

[6] M. Sekula, J. Gaskins, and S. Datta, “Detection of differentially expressed genes in discrete single-cell RNA sequencing data using a hurdle model with correlated random effects,” *Biometrics*, pp. 0–2, 2019.

[7] H. I. H. Chen, Y. Jin, Y. Huang, and Y. Chen, “Detection of high variability in gene expression from single-cell RNA-seq profiling,” *BMC Genomics*, vol. 17, no. Suppl 7, 2016.

[8] P. V Kharchenko, L. Silberstein, and D. T. Scadden, “Bayesian approach to single-cell differential expression analysis.,” *Nat. Methods*, vol. 11, no. 7, pp. 740–2, 2014.

[9] L. Amrhein, K. Harsha, and C. Fuchs, “A mechanistic model for the negative binomial distribution of single-cell mRNA counts,” *bioRxiv*, p. 657619, 2019.

[10] L. de Torrente, S. Zimmerman, D. Taylor, Y. Hasegawa, C. A. Wells, and J. C. Mar, “pathVar: A new method for pathway-based interpretation of gene expression variability,” *PeerJ*, vol. 2017, no. 5, pp. 1–19, 2017.

[11] T. V. de Jong, Y. M. Moshkin, and V. Guryev, “Gene expression variability: the other dimension in transcriptome analysis,” *Physiol. Genomics*, vol. 51, no. 5, pp. 145–158, 2019.

[12] S. H. Yip, P. C. Sham, and J. Wang, “Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data,” *Brief. Bioinform.*, no. November, 2018.

[13] S. H. Yip, P. Wang, J. P. A. Kocher, P. C. Sham, and J. Wang, “Linnorm: improved statistical analysis for single cell RNA-seq expression data,” *Nucleic Acids Res.*, vol. 45, no. 22, p. e179, 2017.

[14] H. Bittersohl and W. Steimer, *Intracellular Concentrations of Immunosuppressants*. Elsevier Inc., 2015.

[15] J. K. Actor, “A Functional Overview of the Immune System and Immune Components,” *Introd. Immunol.*, pp. 1–16, 2019.

[16] J. K. Actor, “Chapter 1 - A Functional Overview of the Immune System and Immune Components,” in *Introductory Immunology*, J. K. Actor, Ed. Amsterdam: Academic Press,

2014, pp. 1–15.

[17] T. P. Monie, *A Snapshot of the Innate Immune System*, no. Section 5. 2017.

[18] J. I. Saldana, “Macrophage Fact Sheet,” *Br. Soc. Immunol.*

[19] T. Hussell, “Helper and Cytotoxic T cells,” p. 3.

[20] M. Wieczorek *et al.*, “Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation,” *Front. Immunol.*, vol. 8, no. MAR, pp. 1–16, 2017.

[21] A. E. Wakil, Z. E. Wang, J. C. Ryan, D. J. Fowell, and R. M. Locksley, “Interferon γ derived from CD4 + T cells is sufficient to mediate T helper cell type 1 development,” *J. Exp. Med.*, vol. 188, no. 9, pp. 1651–1656, 1998.

[22] T. Cd, “CD4 T cells CD8 T cells,” vol. 3, no. d, p. 45.

[23] J. Zhu, “Effector CD4+ T Lymphocytes,” *Ref. Modul. Biomed. Sci.*, vol. 4, 2014.

[24] P. Cruz-Tapias, J. Castiblanco, N. E. Correa, and G. Montoya-Ortíz, *Analysis of Nucleic Acids*. 2013.

[25] J. K. Actor, “The B Lymphocyte: Antibodies and How They Function,” *Introd. Immunol.*, pp. 31–44, 2019.

[26] T. Kurosaki, K. Kometani, and W. Ise, “Memory B cells,” *Nat. Rev. Immunol.*, vol. 15, p. 149, Feb. 2015.

[27] M. . Paridah, A. Moradbak, A. . Mohamed, F. abdulwahab taiwo Owolabi, M. Asniza, and S. H. . Abdul Khalid, “Introductory Chapter: A Brief Overview on Natural Killer Cells,” *Intech*, vol. i, no. tourism, p. 13, 2016.

[28] G. H. Lowell, L. F. Smith, M. S. Artenstein, G. S. Nash, and R. P. Macdermott, “Antibody-dependent cell-mediated antibacterial activity of human mononuclear cells: I.K Lymphocytes and monocytes are effective against meningococci in cooperation with human immune sera,” *J. Exp. Med.*, vol. 150, no. 1, pp. 127–137, 1979.

[29] J. C. Sun and L. L. Lanier, “NK cell development, homeostasis and function: parallels with CD8+ T cells,” *Nat Rev Immunol*, vol. 11, no. 10, pp. 645–657, 2015.

[30] M. Bethesda, “NIH Stem Cell Information Home Page. In Stem Cell Information,” *National Institutes of Health, U.S. Department of Health and Human Services*, 2016. .

[31] T. Stuart *et al.*, “Comprehensive Integration of Single-Cell Data,” *Cell*, vol. 177, no. 7, pp. 1888-1902.e21, 2019.

[32] G. Monaco *et al.*, “RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types,” *Cell Rep.*, vol. 26, no. 6, pp. 1627-1640.e7, 2019.

[33] M. Bots and J. P. Medema, “Granzymes at a glance,” *J. Cell Sci.*, vol. 119, no. 24, pp. 5011–5014, 2006.

[34] P. T. Rudak, J. Choi, and S. M. M. Haeryfar, “Glucocorticoid receptor signaling during prolonged psychological stress compromises the ability of invariant NKT cells to participate in antitumor immune surveillance,” *J. Immunol.*, vol. 200, no. 1 Supplement, pp. 57.48 LP-57.48, May 2018.

[35] A. J. Ridley, “Rho proteins, PI 3-kinases, and monocyte/macrophage motility,” *FEBS Lett.*, vol. 498, no. 2–3, pp. 168–171, 2001.

[36] J. C. Mar *et al.*, “Variance of gene expression identifies altered network constraints in neurological disease,” *PLoS Genet.*, vol. 7, no. 8, 2011.